PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-134214
(43)Date of publication of application : 12.05.2000

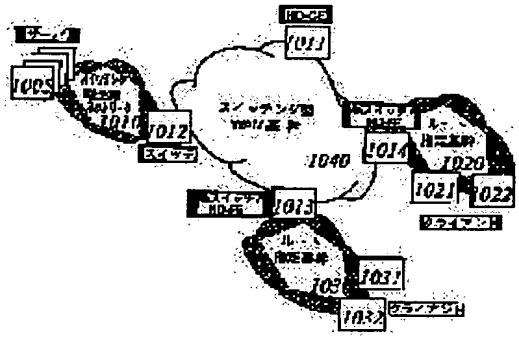
(51)Int.Cl. H04L 12/28
H04L 12/46
H04L 12/56

(21)Application number : 10-301470
(22)Date of filing : 22.10.1998

(71)Applicant : INTERNATL BUSINESS MACH CORP <IBM>
(72)Inventor : GILLENE CRAND
HUNT GUERNEY DOUGLAS HOLLOWAY
ERIC MICHEL LEVY-ABENOORU
DANIEL GEORGES JEAN-MARIE MODYU

(54) DISTRIBUTION TYPE SCALABLE SYSTEM SELECTING SERVER FROM SERVER CLUSTER TO SELECT SWITCHING PATH LEADING TO THE SELECTED SERVER

(57)Abstract:
PROBLEM TO BE SOLVED: To build up a scalable CP/IP service from a switch or a network by transmitting data relating to a request of a client by a transfer engine to a selected server through a switching type connection not through a control engine.
SOLUTION: Network dispatcher transfer engines (FE) 1013, 1014 receive a request of a client and the FE 1013, 1014 transmit a request to a network dispatcher control engine(CE) 1011 in response to the request of the client to select a server and a corresponding switching address from a cluster. The CE 1011 selects the server to send a corresponding switching address to the FE 1013, 1014, and the FE 1013, 1014 transmits data relating to the request of the client to the selected server through a switching type connection relating to the switching address. The switching type connection is not required to pass through the CE 1011.



LEGAL STATUS

[Date of request for examination] 27.07.1999
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number] 3327850

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2000－134214
（P2000－134214A）
(43)公開日　平成12年５月12日(2000.5.12)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｈ０４Ｌ　12/28		Ｈ０４Ｌ　11/20	Ｇ	５Ｋ０３０
12/46		11/00	３１０Ｃ	５Ｋ０３３
12/56		11/20	１０２Ｄ	

審査請求　有　　請求項の数22　ＯＬ　（全 19 頁）

(21)出願番号　特願平10－301470

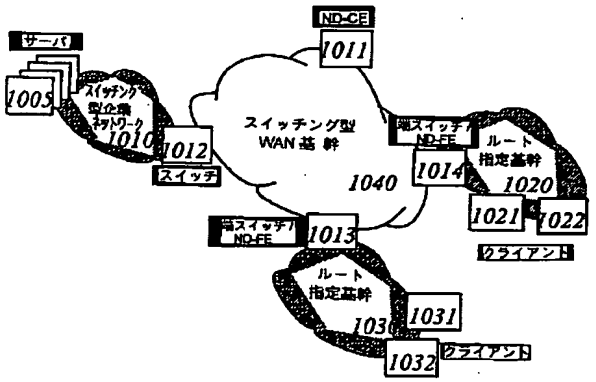
(22)出願日　平成10年10月22日(1998.10.22)

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）
(72)発明者　ジレーヌ・クランド
フランス国06100、ニース、レ・カレブ、アヴニュ・アンリ・デュナン　70
(74)代理人　100086243
弁理士　坂口　博　（外２名）

最終頁に続く

(54)【発明の名称】　サーバ・クラスタからサーバを選択し、選択されたサーバへのスイッチング経路を選択するための分散型スケーラブル装置

(57)【要約】　（修正有）
【課題】ＡＴＭスイッチまたはＡＴＭスイッチング型のネットワークなどのスイッチまたはスイッチング型のネットワークの中からクラスタ・ウエブ・サーバなどのスケール可能なＴＣＰ／ＩＰサービスを構築する。
【解決手段】分散化された大型ＴＣＰルータはＡＴＭスイッチング型のネットワークを用いて構築されうる。スケールされたサービスは単一のサービスとしてクライアントに与えられる。クライアントはスイッチまたはスイッチング型のネットワークに直接または間接に接続されうる。１バージョンは制御エンジン（ＣＥ）および転送エンジン（ＦＥ）と言う２つの要素を含む。ＣＥはサーバに接続を割り当てて、割り当てられたサーバに関する情報を転送し、またＦＥへの接続を割り当てる機能を持つ。ＦＥはＣＥから受け取った割り当てを適用してＴＣＰ接続をスイッチ型ＡＴＭ接続にマップする。終了時に、ＦＥは接続終了事象をＣＥに戻す。



【特許請求の範囲】
【請求項１】スイッチング型のネットワークを含むクライアント－サーバ・システムにおいてサーバのクラスタからあるサーバを選択すると共に選択されたサーバに至るスイッチされる経路を選択するための方法であって、
転送エンジンによりクライアントの要求を受け取るステップと、
クライアントの要求に応答して、クラスタからサーバを選択すると共に対応するスイッチング・アドレスを選択するため、転送エンジンにより要求を制御エンジンに送るルートを指定するステップと、
制御エンジンによりサーバを選択して対応するスイッチング・アドレスを転送エンジンに送るステップと、
転送エンジンにより、クライアントの要求に関連されたデータを、スイッチング・アドレスに関連するスイッチング型接続を通して、制御エンジンを通ることなく選択されたサーバに送るステップと、
を含む方法。
【請求項２】スイッチング・アドレスを転送エンジンに送る前記ステップは、制御エンジンによりサーバ選択基準および転送エンジンがこの基準を使用できる条件を転送エンジンに送り、
転送エンジンにより受け取られる後続のクライアントの要求に対して、
転送エンジンによりこの要求に現存するスイッチング型接続が関連しているか否かを調べ、
現存する接続が存在するならば、転送エンジンにより現存する接続を通してこの要求を送り、
現存する接続が存在しないならば、転送エンジンにより前記基準に基づいて宛先をローカルに選択する、
ステップを含む、
請求項１に記載の方法。
【請求項３】転送エンジンによりスイッチング・アドレスに基づいて、スイッチング型ＡＴＭ接続にＴＣＰ接続をマップするステップを更に含む請求項１に記載の方法。
【請求項４】制御エンジンから遠隔の場所に複数の転送エンジンを配置してこの転送エンジンをスイッチング型ネットワークの端に分散させ、
前記転送エンジンの各々により制御エンジンの指示の下でＴＣＰ接続を分散させるステップ、
を更に含む請求項３に記載の方法。
【請求項５】分散された複数の転送エンジンをスイッチ組織を介して制御エンジンに接続させ、
制御エンジンの指示の下で、分散された各転送エンジンによりスイッチング型接続をマップするステップ、
を更に含む請求項３に記載の方法。
【請求項６】クライアント－サーバ・システムが制御エンジンに障害が生じたときのため一次制御エンジンおよびバックアップ制御エンジンを含んでおり、
前記方法が、
一次制御エンジンの障害を検出するステップと、
前記検出するステップに応答して、バックアップ制御エンジンが一次制御エンジンを引継ぎ、それが新たな一次制御エンジンであることを転送エンジンに通知するステップと、
を含む請求項１に記載の方法。
【請求項７】クライアント－サーバ・システムが複数の転送エンジンを含んでおり、
前記方法が、
クライアントの要求に応答して、構成情報を用いて、一次転送エンジンが障害を起こしたとき選択されうる１つ以上のバックアップ転送エンジンを構成に組み入れるステップと、
一次転送エンジンが障害を起こしたとき、活動中のクライアント接続を中断することなくバックアップ転送エンジンにデータをルート指定するステップと、
を更に含む請求項１に記載の方法。
【請求項８】障害を起こした転送エンジンが回復したことを判定し、回復した転送エンジンを更新するステップと、
新たな要求が回復した転送エンジンにルート指定され、現存する接続のためのパケットをクライアントに中断を与えることなく一次転送エンジンとしての回復した転送エンジンに再ルート指定するようにネットワークを更新するステップと、
を更に含む請求項７に記載の方法。
【請求項９】クライアントがスイッチング型のネットワークに直接取り付けられ、クライアントが転送エンジンを含んでいる請求項１に記載の方法。
【請求項１０】クライアントがインタネットを介してスイッチング型のネットワークに取り付けられる請求項１に記載の方法。
【請求項１１】スイッチング型のネットワークを含むクライアント－サーバ・システムにおいてサーバのクラスタからあるサーバを選択すると共に選択されたサーバに至るスイッチング経路を選択するためのシステムであって、
クライアントの要求を受け取って要求を制御エンジンにルート指定し、クライアントの要求に応答してクラスタからサーバおよび対応するスイッチング・アドレスを選択するための手段と、
前記制御エンジンは、サーバ選択要求に応答してサーバを選択するための制御エンジン手段、および対応するスイッチング・アドレスを転送エンジンに送るための制御エンジン手段を含むことと、
前記転送エンジンは前記スイッチング・アドレスに関連するスイッチング接続を通して制御エンジンを通ることなくクライアントの要求に関連するデータを選択されたサーバに送るための転送エンジン手段と、

を含むシステム。
【請求項１２】前記スイッチング・アドレスを転送エンジンに送るための制御エンジン手段は、
サーバ選択基準および転送エンジンが該基準を使用できる条件を転送エンジンに送るための制御エンジン手段を含み、
その後受け取られるクライアントの要求に対して、転送エンジンが、
この要求に関連する現存するスイッチング接続が存在するか否かを判定するための転送エンジン手段と、
現存する接続が存在する場合、その接続を通して要求を送るための転送エンジン手段と、
現存する接続がない場合、前記基準に基づいて宛先サーバをローカルに選択するための転送エンジン手段と、
を更に含む請求項１１に記載のシステム。
【請求項１３】スイッチング・アドレスに基づいてＴＣＰ接続をスイッチング型ＡＴＭ接続にマップするための転送エンジン手段を更に含む請求項１１に記載のシステム。
【請求項１４】制御エンジンから遠隔の場所にあるスイッチング型ネットワークの端に分散された複数の転送エンジンを含み、
前記転送エンジンの各々が制御エンジンの指示の下でＴＣＰ接続を分散させることを特徴とする請求項１３に記載のシステム。
【請求項１５】分散された複数の転送エンジンをスイッチ組織を介して制御エンジンに接続させ、
制御エンジンの指示の下で、分散された各転送エンジンによりスイッチング型接続をマップする、
ことを特徴とする請求項１３に記載のシステム。
【請求項１６】制御エンジンに障害が生じたときのために一次制御エンジンおよびバックアップ制御エンジンを含んでおり、
一次制御エンジンの障害を検出する手段と、
前記検出する手段に応答して、バックアップ制御エンジンが一次制御エンジンを引継ぎ、それが新たな一次制御エンジンであることを転送エンジンに通知するためのバックアップ制御エンジン手段と、
を更に含む請求項１１に記載のシステム。
【請求項１７】複数の転送エンジンを含んでおり、
クライアントの要求に応答して、ネットワークで得られる構成情報を用いて、一次転送エンジンが障害を起こしたとき選択されうる１つ以上のバックアップ転送エンジンを構成に組み入れるための手段と、
一次転送エンジンが障害を起こしたとき、活動中のクライアント接続を中断することなくバックアップ転送エンジンにデータをルート指定するための手段と、
を更に含む請求項１１に記載のシステム。
【請求項１８】障害を起こした転送エンジンが回復したことを判定し、回復した転送エンジンを更新するための手段と、
新たな要求が回復した転送エンジンにルート指定され、現存する接続のためのパケットをクライアントに中断を与えることなく一次転送エンジンとしての回復した転送エンジンに再ルート指定するようにネットワークを更新するための手段と、
を更に含む請求項１７に記載のシステム。
【請求項１９】クライアントがスイッチング型のネットワークに直接取り付けられ、クライアントが転送エンジンを含んでいる請求項１１に記載のシステム。
【請求項２０】クライアントがインタネットを介してスイッチング型のネットワークに取り付けられる請求項１１に記載のシステム。
【請求項２１】サーバがルート指定されるネットワークに取り付けられる請求項１１に記載のシステム。
【請求項２２】転送エンジン、制御エンジン、およびスイッチ組織が単一の装置に共存する請求項１５に記載のシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】インタネット・エンジニアリング・タスク・フォース（ＩＥＴＦ）のＮＢＭＡ（non-broadcast multiple access）でのインタネットワーキング（Internetworking over NBMA－ＩＯＮ）のワーキング・グループはインタネット・プロトコル（ＩＰ）スイッチングに対する３つの異なる提案を現在検討中である。これらのアーキテクチャは２つの方法により要約することができる。即ち先ず、イプシロンスイッチング方法は非同期転送モード（ＡＴＭ）接続をインタネット・プロトコル・フローに関連づけ、次のものは接続を出口ルータのルートに関連づける。
【０００２】
【従来の技術】ＡＴＭはこの分野で良く知られている。概観すると、ＡＴＭは歴史的にはＢ－ＩＳＤＮ（Broadband Integrated Services Digital Network）の開発に基づいている。ＡＴＭはＢ－ＩＳＤＮの伝送モードとして選ばれたパケットを多重化しスイッチングする方法である。ＡＴＭ、即ち高速ディジタル伝送、はパケット・スイッチング技術を用いるが、「非同期」伝送とは関係がない。（例えば、PDH,Broadband ISDN, ATM and All That:A Guide to Modern WAN Networking, and How It Evolved,Paul Reilly著、Silicon Graphics Inc. April 4, 1994発行を参照されたい）。ＡＴＭパケットはセルと呼ばれ、各セルは５バイトのヘッダおよび４８バイトのデータを有する。ＡＴＭパケット・スイッチングは、ＡＴＭパケットが仮想的経路および仮想的回路と呼ばれる予め設定されたルートを辿ると言う点において従来のパケット・スイッチングと異なる。ＡＴＭはいかなる特定の物理的伝送媒体にも依存しないが、伝送媒体が主として光ファイバであるときにはエラー率および損失率が

非常に少なく、従って再伝送は行われない。"Asynchronous Transfer Mode Tutorial", Northern Telecom, http://www.webproforum.com/nortel2/index.html, (6/10/98)を参照されたい。

【０００３】トランスミッション・コントロール・プロトコル／インタネット・プロトコル（ＴＣＰ／ＩＰ）およびＡＴＭにおけるＴＣＰ／ＩＰの使用はこの分野で良く知られている。D.E. Comer著、"Internetworking with TCP/IP:Principles, Protocols, and Architecture", Englewood Cliffs, N.J.,Prentice Hall発行（1988）を参照されたい。トランスミッション・コントロール・プロトコル（ＴＣＰ）スイッチングは異なるＡＴＭルータの間で接続を割り当てることによって、例えば予め定義された仮想経路インディケータ／仮想チャネル・インディケータ（ＶＰＩ／ＶＣＩ）を用いて働くが、この方法は接続当たり交換される所与の量のパケットが効率的であることを必要とする。取りうる別の処理方法はイプシロンＩＰスイッチング方法を用いることである。

【０００４】ワールドワイド・ウエブ上のトラフィックは特に評判の良い（ホット）所では指数的に増大している。従ってスケーラブルなウエブ・サーバを提供することが重要である（例えば、Goldszmidt, G.および Hunt, G.著、"Net Dispatcher a TCP Connection Router" IBM Research Report, 1997,および Dias, D.M., Kish, W., Mukherjee, R., Tewari, R.著、"A Scalable and Highly Available Web Server", Proc. 41st IEEE Computer Society Intl. Conf. (COMPCON), 1996, Technologies for the Information Superhighway, pp85-92, Feb. 1996 を参照されたい。スケーラブル・ウエブ・サーバにおいて負荷バランスを与える公知の一方法はネットワーク・ディスパッチャを用いることである（例えば、米国特許第5,371,852号、およびAttanasio, Clement R, Smith, Stephen E.著、"A VirtualMulti-Processor Implemented by an Encapsulated Cluster of Loosely Coupled Computers", IBM Research Report RC 18442 (1992)を参照されたい）。ここではネットワーク・ディスパッチャ（ＮＤ）のアドレスだけがクライアントに与えられ、ネットワーク・ディスパッチャがクラスタ（いわゆる仮想的カプセル化クラスターＶＥＣ）内のノードの間で、ラウンド・ロビン式に、またはノードの負荷に基づいて、入来要求を分配する。継続中の米国特許出願番号08/861,749には一般化されたネットワーク・ディスパッチャの例が開示されており、これはネットワーク相互間全体の任意の場所に所在するノードへのルート決定を可能にする。

【０００５】インタネットの基幹ネットワークはスイッチング型ＡＴＭインフラストラクチャに移行しつつある。同時に、非常に大きなサーバ（メイン・フレーム、メイン・フレーム・クラスタ、またはその他のタイプのクラスタの如何を問わず）が、帯域幅および要求時スループットの画期的な成長を扱うためＡＴＭリンクを介してインタネット基幹に接続されようとしている。

【０００６】この意味において、ＩＥＴＦはＡＴＭスイッチの単純、高速、効率的な処理能力の利点を活用するための種々の代替案を検討中である。種々の代替案に共通するものは、ＩＰヘッダに基づくルート決定をＡＴＭヘッダに基づくスイッチング決定と置き換えてすべての中間のホップ（クライアントおよびサーバ以外のすべてのホップ）処理を単純化するための動的な方式である。このことは究極的にはエンドポイント（つまり、クライアントおよびサーバ）だけがＩＰパケット（ＩＰレイヤー、ＴＣＰレイヤーなど）を処理し、エンドポイント間の経路上のその他のすべてのホップがＡＴＭパケットをスイッチすることを意味する。代替案の或るものはいわゆる「ショートカット」方法も検討しており、これは物理的接続性が許容するときには中間のホップの幾つかをバイパスする機構である。インタネット業界で検討されている解決策には、ネクスト・ホップ・レゾリューション・プロトコル（ＮＨＲＰ）、イプシロンＩＰスイッチング・プロトコル（ＩＦＭＰおよびＧＳＭＰ）、タグ・スイッチング、およびＩＢＭのアグリゲート・ルート・ベース・ＩＰスイッチ（ＡＲＩＳ）がある。ＮＨＲＰについては、"Next Hop Resolution Protocol (NHRP)", The Internet Society, Network Working Group, RFC 2332 (1998)を参照されたい。

【０００７】全面的にまたは部分的であってもスイッチング型のネットワークにおいては、サーバのクラスタに対し従来型のフロント・エンド（ネットワーク・ディスパッチャなど）を稼働させるホップはＩＥＴＦで検討されている革新的なアプローチ全体と相容れないものとなる。ＩＰおよびＴＣＰ分野を調べることが必要であり、その他のどのようなホップもルート決定を行うのにＩＰを考慮するのを避けようとすることになるであろう。

【０００８】

【発明が解決しようとする課題】上述の必要性に従って、本発明はサーバのクラスタに対するフロント・エンドにスイッチング能力を与えて、パケットがサーバまで、また、サーバに戻るように、またはクライアントに最も近いスイッチへと、スイッチされるようにする特徴を有する。

【０００９】

【課題を解決するための手段】本発明の１態様は１つの要素、すなわち、制御エンジン（ＣＥ）および転送エンジン（ＦＥ）を含む。制御エンジンはサーバへの接続を割り当てて、割り当てられたサーバに関する情報を伝送し、そして転送エンジンに接続する機能を有する。各転送エンジンは制御エンジンから受け取った割り当てを適用してスイッチング型ＡＴＭ接続へのＴＣＰ接続をマップする。転送エンジンは終了時に接続終了事象を制御エンジンに戻す。

【００１０】スイッチング型のネットワークを含むクライアントーサーバ・システムにおいてサーバのクラスタからサーバを選択し、選択されたサーバに至るスイッチされた経路を選択する方法の例は以下のステップを含む。即ち、転送エンジン（ＦＥ）がクライアントの要求を受け取るステップと、クライアントの要求に応答してＦＥが制御エンジン（ＣＥ）に要求を送ってクラスタからサーバおよび対応するスイッチング・アドレスを選択するステップと、ＣＥがサーバを選択して対応するスイッチング・アドレスをＦＥに送り、ＦＥがクライアントの要求に関連するデータをスイッチング・アドレスに関連するスイッチング型接続を通して選択されたサーバに送るステップとが含まれる。ここでスイッチング型接続はＣＥを通る必要はない。

【００１１】ＣＥがスイッチング・アドレスをＦＥに送るステップは更に以下のステップを含む。即ち、ＣＥがサーバ選択基準およびＦＥがその基準を使用することができる条件をＦＥに送るステップと、ＦＥにより受け取られたその後のクライアント要求に対して、現存するスイッチされた接続がこの要求に関連しているかどうかをＦＥが判定するステップと、現存する接続があるならば、現存するスイッチされた接続を通してＦＥがこの要求を送るステップと、現存するスイッチされた接続がないならばＦＥが基準に基づいて宛先をローカルに選択するステップと、が含まれる。

【００１２】本発明はその他の特徴を有しており、これは中央に集中されたクラスタ・フロント・エンド（クラスタ・サーバまたはディスパッチャとも呼ばれる）を通るパケットの経由を最小にし、ルート決定機能の幾分かをスイッチング型のネットワークの端に分配することによって隘路が生じる可能性を都合良く低減する。また、この方法はディスパッチャに２ホップ以上離れたサーバを管理させ、また伝送処理を分配することによってシステム全体の頑強さおよび性能を増大させる。例えば、ＦＥはスイッチング・アドレスに基づいてＴＣＰ接続をスイッチング型ＡＴＭ接続にマップすることができる。ＣＥから遠くに離れた複数のＦＥを設けてＦＥがスイッチング型のネットワークの端に分布され、各ＦＥがＣＥの指図に従ってＴＣＰ接続を分配するようにすることもできる。別の例では、複数の分散されたＦＥがスイッチ組織を介してＣＥに接続されることも可能であり、分散された各ＦＥはＣＥの指図に従ってスイッチング型接続をマップする。

【００１３】本発明の１実施例は現存するＴＣＰ接続ルータのすべての能力を含んでいる。これには利用可能度が高いことおよびフォールト・トレランスが含まれる。例えば、１９９７年９月１５日出願、継続中の米国特許出願（出願人整理番号Y0997-232）を参照されたい。別の実施例は現存するＴＣＰ接続型のルータの、サーバ・クラスタに加わる負荷をバランスさせるためのフィードバック機能の幾つかを含んでいる。

【００１４】本発明によるフォールト・トレランスの特徴を含む方法の１例は一次ＣＥと、この一次ＣＥが故障したときのためのバックアップＣＥとを含んでいる。この方法は、一次ＣＥの障害を検出するステップと、障害の検出に応答してバックアップＣＥが一次ＣＥを引き継いでＦＥにそれが新しい一次ＣＥであることを通知するステップとを含む。

【００１５】複数のＦＥを含む別の例はクライアントの要求に応答してネットワークで入手できるコンフィギュレーション情報を用いて一次ＦＥが障害を起こしたときに選択されるバックアップＦＥを１つ以上構成に組み入れるステップと、一次ＦＥが障害を起こしたとき活動中のクライアント接続を中断することなくバックアップＦＥにデータを送るステップとを含む。追加のステップとして、障害を起こしたＦＥが回復したことを判定するステップと、回復したＦＥを現状に更新するステップと、ネットワークを更新して新たな要求が回復したＦＥに送られるようにするステップと、クライアントを中断することなく現存する接続のパケットを一次ＦＥとして回復したＦＥに経路変更するステップとが含まれても良い。

【００１６】本発明は取り付けられた（直接または間接に）サーバによって提供されるサービスをスケールする汎用装置を開発するためのシステムおよび方法を含む。この汎用装置は強化されたスケール能力を与えるためのスイッチを利用する。この装置の１例において、分散大型ＴＣＰルータがＡＴＭスイッチング型のネットワークを用いて構築される。スケールされたサービスはクライアントに対する単一のサービスとして提供されうる。これらのサービスはスイッチ組織に直接または遠隔地から取り付けられ得る。

【００１７】幾つかの利点として、次のものが挙げられる。

・インタネット・サービスをスケールする現在のいかなる手法と比べても最高の容量およびスループットを有する。

・この手法はフォールト・トレランスおよび高い利用可能度を有する。

・この手法はサーバおよびクラスタ・サーバが同一サブネットに共存することを強制すると言うような制約を持たない。サーバはＡＴＭ組織に直接取り付けられまたはルート経由のネットワークを介して取り付けられ得る。

・スイッチ組織はネットワーク内にあって良く、または高度にスケール可能な並列コンピュータであっても良く、或いはその他任意のアプリケーション（電話など）であって良い。

【００１８】

【発明の実施の形態】図１はスイッチング技術を含むスイッチング型のネットワークに配置された本発明の例を示す。一般に、スイッチング・ワイド・エアリア・ネッ

トワークの基幹１０４０は、フレーム・リレー、ＡＴＭ、またはＸ．２５を含む（これらに限られるものではないが）任意のスイッチ技術により具体化された従来型のＷＡＮ（Wide Area Network）である。同様にスイッチされる企業ネットワーク１０１０は、フレーム・リレー、ＡＴＭ、またはＸ．２５等の任意のスイッチ技術により具体化された従来型の企業ネットワークである。従来通り、基幹ネットワークはインタネットに対する中央の相互接続を作る任意のネットワークである。全国的基幹は通常はＷＡＮであり、企業基幹は通常はＬＡＮ（Local Area Network）またはＬＡＮの組合せである。

【００１９】本発明は周知のネットワーク・ディスパッチャ（ＮＤ）を含むことが望ましい。これはＴＣＰ接続のソフトウエア・ルータであり、複数のＴＣＰサーバ間で負荷のバランスをも支援する。しかしながら、任意のＴＣＰ接続ルータや負荷バランス論理が本発明に採用できることは当業者にとって明らかであろう。図に示されたように、本発明は２つの主体であるネットワーク・ディスパッチャ制御エンジン（ＮＤ－ＣＥ）１０１１およびネットワーク・ディスパッチャ転送エンジン（ＮＤ－ＦＥ）１０１３、１０１４を含む。ＮＤ－ＣＥ１０１１はサーバ１００５への接続を割り当てて、そのサーバに関する情報を送り、またＮＤ－ＦＥ１０１３、１０１４への接続を割り当てる。ＮＤ－ＦＥ１０１３、１０１４の各々はＮＤ－ＣＥ１０１１から受け取った割り当てを適用してスイッチング（例えばＡＴＭ）接続へのＴＣＰ接続をマップする。ＮＤ－ＦＥは最終的には接続終了事象をＮＤ－ＣＥ１０１１に戻す。この方法がＮＤ－ＣＥ１０１１とＮＤ－ＦＥ１０１３、１０１４との間で情報を伝播させるのに用いられるプロトコルに無関係であると共に、これらの機能の物理的所在にも無関係であることは当業者に明らかであろう。

【００２０】転送エンジン（ＦＥ）は次の２つの極端に至るまでネットワークのどこにあっても良い。

１．クライアント側：ＴＣＰ接続がクライアントからサーバへ全体的にスイッチされるようにする。この手法は最も効率的であり、クライアントがスイッチング型のネットワーク（１０４０または１０１０）に直接接続されることを必要とする（より詳細な例が図４を参照して後で説明される）。しかしながら一般にクライアント（１０３１、１０３２、１０２１、１０２２）はスイッチング型のネットワークを介して直接に接続される必要はなく、これらは任意の技術を用いてインタネットに接続されても良い。

【００２１】２．制御エンジンと同じ場所：クラスタとなったＮＤ－ＦＥおよびＮＤ－ＣＥの組は従来型のスイッチ組織または企業スイッチング型のネットワークを介して接続されうる。このやり方はクライアントおよびＷＡＮの基幹を変更しないと言う利点を持つ。しかしながら、ＷＡＮ基幹がスイッチング型のネットワーク１０４０であるならば、この手法はスイッチング型ＷＡＮ基幹１０４０を完全には利用しないことになる。それでもこの手法は転送エンジン１０１４の分散を可能にすると共に企業スイッチング型のネットワーク（併合されたＮＤ－ＦＥ／ＮＤ－ＣＥとサーバとの間）の利点を依然活用する。併合されたＮＤ－ＦＥ／ＮＤ－ＣＥの例は図３を参照して後で説明される。

【００２２】好適なＴＣＰ接続ルータの実施例では、ＦＥは「実行者」プロセスを含み、ＣＥは「実行者」および「管理者」プロセスを含む。これらは例えばＮＤから適応されたものである。「実行者」は高速ＩＰパケット転送を支援するＯＳカーネルの拡張であって良く、また「管理者」は「実行者」を制御するユーザ・レベルのプロセスであって良い。この新しい実施態様はその「実行者」を有するＦＥがスイッチまたはスイッチング型のネットワークの端に分散されることを可能にして性能および頑強性の改善をもたらす。ＦＥはＣＥの指示の下で同期的または非同期的に接続を分散する。ホストは従来から通信していたのと同じ方法でＣＥと通信することもできる。一次と指定されるＣＥおよび二次と指定されるもう１つのＣＥと言う２つのＣＥをネットワークに持つことによって高い利用可能性およびフォールト・トレランスが得られる。２つのＣＥ間の通信は公知の方法を用いることができる。ＦＥが分散されているので相違点としては一次とバックアップとの間で追加の状態が転送されねばならないことが挙げられる。

【００２３】従来型のネットワーク・ディスパッチャまたはルータとは異なり、本発明はトポロジ、速度、リンク速度、およびクライアントの要求をサーバに伝えるためにスイッチング型のネットワークに備えられるその他の情報を用いることができる。例えば、ＡＴＭなどのスイッチング型のネットワークではＰＮＮＩ（PrivateNetwork to Network Interface）などの経路指定プロトコルがスイッチに大量の情報を分配し、これがＮＤ－ＣＥやＮＤ－ＦＥにより用いられて適当なサーバに至る適当なルートを選択する。このようなルート決定プロトコルを利用する場合、ＮＤ－ＣＥおよびＮＤ－ＦＥは、サーバ選択を改善するのに用いられるその他のスイッチ、それらの間のリンク、および装置に関するおそらくは多様な関連情報にアクセスする。このような有用な情報の例には、端から端までの遅延（サーバに至るまで）ジター（サーバの至るまでの遅延の変動）、サーバへの、およびサーバからのスループット（平均、ピーク、バースト度）、および伝播遅延が含まれる。これは構成可能な計測量が最適なサーバを選択するために基準となることを可能にする。この決定はＮＤ－ＣＥにおいてなされ、ＮＤ－ＣＥは、重みおよびＮＤ－ＦＥがその基準を用いることができる条件などの決定基準をＮＤ－ＦＥ１０１３、１０１４に送る。このようにしてＮＤ－ＦＥはそれが作った現存するスイッチ接続を用いてサーバに接続要

求を独立的に割り当てることができる。ND－CEのクライアントであるND－FEの部分がこの情報を優先的に受け取ってND－FEにおけるその使用を指示することになることは当業者にとって明らかであろう。

【0024】図1に戻ると、1つ以上のクライアント1021、1022がSWAN（Switched Wide Area Network）基幹1040にルートされた基幹1020（ルートされたネットワークとも呼ばれる）およびES／ND－FE（Edge Switch/Network Dispatcher Forwarding Engine）1014を介して接続されている。クライアント1031、1032はルートされたネットワーク1030およびES／ND－FE1013を介してSWAN基幹1040に接続されている。ルートされた基幹1020、1030は、クライアントの要求を端スイッチ1013または1014に送ることのできる任意の基幹として広義に定義される。スイッチされる企業ネットワーク1010もスイッチ1012を介してSWAN基幹1040に接続される。ネットワーク・ディスパッチャ制御エンジンND－CE1011もSWAN基幹1040に接続される。サーバのクラスタ1005はスイッチされる企業ネットワーク1010に接続される。

【0025】ネットワーク1040および1010がATMネットワークである場合にはNHRP（Next Hop Resolution Protocol）が本発明に従って使用されうる。NHRPは同じサブネットに属さない端点間にいわゆるショートカット接続を作り、中間のすべてのNBMA（non-broadcast multiple access）を取り付けられたルータをバイパスすることを可能にする。標準的なNHRP要素はNHRPクライアント（NHC）およびNHRPサーバ（NHS）を含む。この場合、ND－FEおよびND－CEはNHC＋＋（以下に説明する追加の機能を有する標準的なNHC）と呼ばれる変更されたNHRPを用いることができる。一実施例において、ND－CE1011はNHS＋＋（以下に説明する追加の機能を有する標準的なNHS）と呼ばれる変更されたNHRPを含んでいる。

【0026】サーバのクラスタ1005には幾つかの目標となりうるものがあるが、クライアントはただ1つの目標IPアドレス（仮想的なカプセル化されたクラスタ（VEC）アドレス）しか見ていないので、NHRPの具現化には特別の拡張が加えられなければならない。従ってNHC＋＋およびNHS＋＋機能（以下に述べる）がこの場合設けられなければならない。必要とされる特定の装置の数は顧客のどのような構成でも支援できるように最小にされるべきである。ショートカット接続の利点を最大限に利用するために、好適な実施例はNHC＋＋クライアントをSWAN1040への入り口ルータに置き、NHS＋＋機能を少なくともND－CE1011に置く。WAN入り口NHC＋＋からND－CE／NHS＋＋への経路にあるルータに必要とされるすべてのことはNHSの支援である。これと同様に、ND－CE／NHS＋＋から目標サーバのNHC＋＋への経路にあるルータに必要とされるすべてのことはNHSの支援である（図7を参照してこの後述べられるように）。企業ルータ（WAN出口1）はNHS＋＋機能を持つ必要はない。

【0027】図2は本発明に従ってクライアントとサーバとの間でスイッチング経路を設定するための論理フローの例を示す。ここに示されたように、クライアント1021はサーバのクラスタ1005からTCPサービスを得るための要求2035を出す。この要求2035は、スイッチング型ワイド・エリア・ネットワーク基幹1040の境界にある端スイッチ1014に最終的に到達する。この端スイッチ1014はネットワーク・ディスパッチャ転送エンジン（ND－FE）2023も含んでいる。ND－FE2023は、要求が現存する接続の一部であるか否かを調べるために標準的なテーブル・ルックアップを行う。要求が現存する接続の一部であるならば、ND－FE2023は対応するスイッチング型接続を取り出して、現存するスイッチング型接続上のクライアント要求をクラスタ1005のサーバにそのまま送る（スイッチング型のネットワーク1040、スイッチ1012、およびスイッチング型企業ネットワーク1010を介して）。既に存在する接続がないならば、ND－FE2023はクライアント要求2025をネットワーク・ディスパッチャ制御エンジンND－CE1011に送る。ND－CE1011はクラスタ1005においてサーバを選択し、選択されたサーバのスイッチング・アドレスをND－FE2023に戻す。

【0028】これに加えて、ND－CE1011はその決定基準（重みなどの）およびND－FE2023がこの基準を用いることのできる条件をND－FE2023に送ることができる。ND－FE2023は、それが作った現存するスイッチ接続を用いてサーバ1005に接続要求を独立して割り当てるためにこの基準を用いることができる。また、ND－CE1011は、サーバへのスイッチング経路をそれがアイドルになった後どのくらい長く維持するかについての情報をND－FE2023に与えることが望ましい。この追加の情報は同じフロー（2026）を用いて、または別途（2029）送ることができる。

【0029】接続設定に戻ると、ND－CE1011は、スイッチング型のネットワーク1040、スイッチ1012、およびスイッチング型企業ネットワーク1010を介してクラスタのサーバ1005に最初のクライアント要求を送る（2015）。ND－FE2023が選ばれたサーバのスイッチング・アドレスを受け取ると、それは選択されたサーバへのスイッチング型接続2028を作ることになる（2027）。スイッチ接続が既に存在するならば（2028）、新しく接続を作る代

わりに（２０２７）現存する（２０２８）接続をそれが再使用することが望ましい。接続が作られた後（新たな接続かまたは現存する接続）、ＮＤ－ＦＥ２０２３は、スイッチング型のネットワーク１０４０、スイッチ１０１２、およびスイッチング型のネットワーク１０１０を介して、作られたスイッチ接続２０２８上のクライアント接続のすべての後続パケット２０３６をサーバ１００５に送ることになる。

【００３０】クライアント１０２１がサーバ１００５への接続を終了すると、ＮＤ－ＦＥ２０２３はその接続を除去としてマークし、接続終了パケットおよびその接続に対するすべての後続パケットをＮＤ－ＣＥ１０１１に送るか（２０２５）、または接続終了パケットをサーバ１００５に送り、接続が終了した後、ＮＤ－ＣＥ１０１１に接続が終了したことを別途通知する（２０２９）。接続終了パケットおよび後続パケットがＮＤ－ＣＥ１０１１に送られようとするとき（２０２５）、ＮＤ－ＣＥ１０１１はこの接続を除去としてマークし、パケットを関連するサーバに送る（２０１５）。接続が或る時間の間アイドルになっているとＮＤ－ＣＥ１０１１はこの接続を除去するが、これは構成可能であることが望ましい。ＮＤ－ＦＥ２０２３がＮＤ－ＣＥ１０１１に接続終了を別途通知すると（２０２９）、ＮＤ－ＣＥ１０１１はその接続テーブルから単純にその接続を除去するだけとなる。接続が終了すると、端スイッチ１０１４とサーバ１００５との間のスイッチング型接続は、同じＮＤ－ＦＥ２０２３から同じサーバに向けられる追加の接続がそれを再使用することができるように維持される。どのクライアント接続もそのサーバを要求することなく或る時間が経過した後、対応するスイッチ接続は除去されても良い。

【００３１】図３は、ＮＤ－ＦＥ、ＮＤ－ＣＥ、端スイッチ、企業基幹へのスイッチが同じ物理的ボックスまたは装置に取り込まれた例を示す。図に示すように、クライアント３０８０、３０８１、３０８２はルート指定されたネットワーク３１１０、ＮＤ－ＣＥ－ＦＥスイッチ３０１０、およびスイッチング型のネットワーク３１００を介してサーバ３０９０、３０９１、３０９２のクラスタにアクセスすることができる。クライアント３０８０からの最初の要求３００５はＮＤ－ＦＥ３０２０に到達する。ＮＤ－ＦＥ３０２０はその接続テーブルでのルックアップが不首尾となった後、最初のクライアント要求３００５をＮＤ－ＣＥ３０４０に送る（３０２５）。ＮＤ－ＣＥ３０４０はクラスタからの選択されたサーバ３０９０のスイッチング・アドレスで以てＮＤ－ＦＥ３０２０に応答する。ＮＤ－ＣＥ３０４０はまた最初のクライアント要求３００５を選択されたサーバ３０９０にも送る（３０３５）。同じクライアント接続上でクライアント３０８０により出されるすべての後続パケット３０４５はクライアント３０８０からＮＤ－ＦＥ３０２０にルートされ、そしてＮＤ－ＦＥ３０２０から選択されたサーバ３０９０にスイッチされる（３０５５）。終了のフローは図２に述べたのと同じである。

【００３２】図４はＮＤ－ＦＥ１０１４および図１のクライアント１０２１が併合されてクライアント／ＮＤ－ＦＥ４４２０となった例を示す。図に示されたように、クライアント／ＮＤ－ＦＥ４４２０が新たな要求を出すと、それはＳＷＡＮ１０４０を介して直接（４４２０）ＮＤ－ＣＥ１０１１に行き、ＮＤ－ＣＥ１０１１はクラスタの選択されたサーバ１００５のスイッチング・アドレスをクライアント／ＮＤ－ＦＥ４４２０に戻す（４４１５）。ＮＤ－ＣＥ１０１１は選択されたサーバ１００５にも要求を送る（４４２５）。この最初の交換の後、この接続についてのすべてのトラフィックはスイッチされる基幹１０４０を介してクライアント／ＮＤ－ＦＥ４４２０と選択されたサーバ１００５との間でスイッチされる（４４４５）。終了のフローは図２について述べたものと同じである。

【００３３】図５はサーバを選択し、そのサーバへの経路を選択または設定するためにＮＤ－ＦＥで用いられる論理フローの例を示す。ステップ５０１０において、ＮＤ－ＦＥ２０２３（図２）はクライアント要求を受け取る。ステップ５０３０において、ＮＤ－ＦＥ２０２３は、この要求が属する現存する接続があるか否かを調べるためにテーブル・ルックアップを行う。現存する接続があるならば、ステップ５１７０においてそれは関連のスイッチング型接続を介してこの要求を単に送るだけである。ステップ５０３０において現存する接続がなければ、ステップ５０６０においてＮＤ－ＦＥ２０２３は、宛先サーバをローカルで選択することができるか否か、またはＮＤ－ＣＥ１０１１まで行かなければならないかどうかを調べる（図２）。この決定は前のフロー（２０２９－図２で述べた）でＮＤ－ＣＥ１０１１により与えられた構成可能な機能およびデータを用いてなされるのが望ましい。ＮＤ－ＦＥ２０２３がサーバをローカルで選択することができる場合には、ステップ５１００においてそれはサーバを選択する。ステップ５０６０においてＮＤ－ＦＥ２０２３がサーバをローカルで選択することができない場合には、ステップ５０９０においてＮＤ－ＦＥ２０２３はＮＤ－ＣＥ１０１１を調べてサーバ選択および対応するスイッチング・アドレスを得る（図２の２０２５）。サーバがローカルで、またはＮＤ－ＣＥ１０１１により選択されると、処理はステップ５１３０に続く。ステップ５１３０において、ＮＤ－ＦＥ２０２３は、選択されたサーバへの現存するスイッチング型接続が存在するか否かを調べる。スイッチング型接続があるならば、ステップ５１７０においてそれは現存するスイッチング型接続を通してこの要求を送る（図２の２０２８）。サーバへの現存するスイッチング型接続が存在しないならば、ステップ５１６０においてＮＤ－ＦＥは

選択されたサーバへのスイッチング型接続を設定する。これがなされると、ステップ５１７０においてそれは新たなスイッチされる（図2の２０２７）接続を通してクライアント要求を送る。

【００３４】図6ないし１２は、割り当てられたサーバをＮＤ－ＦＥに知らせてそれがスイッチされる（ＡＴＭ）接続にＴＣＰ接続をマップすることができるようにするためＮＤ－ＣＥとＮＤ－ＦＥとの間でＮＨＲＰプロトコルを用いる本発明の例を示す。本発明の一部はＮＤ－ＦＥ，ＮＤ－ＣＥ、およびサーバの間の経路に関し、またサーバ上でＮＨＲＰの公知の特徴を用いる。本発明がその他のタイプのスイッチ、またはスイッチング型のネットワークにも直ちに実施できることは当業者にとって明らかであろう。好適な実施例に用いられるすべてのフローは標準的なＮＨＲＰフローである。ＮＨＲＰは拡張フィールドを許容するが、これは追加の機能を取り入れるために用いるのに好都合である。従って、フローに関するすべてのエラーはこの分野で知られた技術を用いて具合良く処理される。

【００３５】図6は本発明を取り入れたネットワーク・トポロジの例を示す。ＴＣＰクライアント（１０１）は、サーバ１４１、および１４２を含むサーバのクラスタのサービスを用いることを必要とするＩＰホストである。クライアントはサーバの1つ（１４１または１４２）上のアプリケーションとＴＣＰ接続を作らなければならない。この例において、クラスタのＩＰアドレスはＩＰ＿ＳＣであるものと仮定される。ＴＣＰクライアントはクラスタのＩＰアドレスＩＰ＿ＳＣおよびＴＣＰポート番号を知っているだけである。サーバはＮＢＭＡ（Non-Broadcast Multiple Access）ネットワーク（１６２）内に置かれる。この図および以降の図において、ＮＢＭＡネットワークにおけるスイッチング型接続はＮＢＭＡ接続と呼ばれることになる。ここで判るように、ネットワーク１６２はＡＴＭネットワークであるが、本発明がその他のタイプのスイッチング型のネットワークにおいても実施できることは当業者にとって明らかであろう。

【００３６】普通は、クライアント１０１によってサーバの1つ１４１または１４２に送られるＩＰデータグラムは点線で示された経路に通常は従う。ルート指定されたこの経路は幾つかの中間のルータ、つまりサーバ１４１については１１１、１３１、１３２、１２１および１３４、サーバ１４２については１１１、１３１、１３２、１２１および１３５と交差する。

【００３７】本発明によれば、ネットワーク１６２にわたってＴＣＰ接続の間ショートカットＡＴＭ接続が作られて中間のホップ数が最小になるようにされる。

【００３８】中間のルータの内の2つ１１１および１２１が本発明に従って用いられている。ルータ１１１はＮＤ－ＦＥを含んでいる。それはサーバとのショートカット接続を作る。それはまた、データグラムをＴＣＰクライアントからショートカット接続に送る。ルータ１２１はＮＤ－ＣＥを含んでいる。それは各新たなＴＣＰ接続毎にサーバを割り当てる。

【００３９】以下のフローはＩＥＴＦ（Internet Engineering Task Force）によって標準化されようとしているＮＨＲＰからのものである。ＮＨＲＰ要素はＮＨＲＰクライアント（ＮＨＣ）およびＮＨＲＰサーバ（ＮＨＳ）を含むが、これらは共にこの分野では知られている。ＮＤ－ＦＥ（１１１）およびＮＤ－ＣＥ（１２１）はＮＨＣ＋＋（以下に述べる追加的機能を有する標準的なＮＨＣ）と呼ばれる変更されたＮＨＲＰクライアントを用いる。ルータ１３３もＮＨＳ＋＋（以下に述べる追加的機能を有する標準的なＮＨＳ）と呼ばれる変更されたＮＨＲＰサーバを含んでいる。必要とされる特定の装置の数はどのような構成をも支援できるように最小でなければならない。ＷＡＮ入り口ＮＨＣ＋＋１１１からＮＨＳ＋＋１３３間での経路にあるルータ１３１、１３２に必要とされるすべてのことはＮＨＳの支援である。

【００４０】ショートカットの利点を十分に活用するための好適な実施例はＮＨＣ＋＋クライアントを、少なくともＮＤ－ＣＥ１２１においてＷＡＮおよびＮＨＳ＋＋機能に至る入り口ルータ１１１に置くことである（ルータ１３３の代わりに）。この場合、ルータ１３３はＮＨＳも支援する。

【００４１】これと同様に、ＮＤ－ＣＥにサービスを与えるＮＨＳ＋＋（１３３）から目標サーバ１４１、１４２への経路にあるすべてのルータ（図6には何も示されていないが、ルータ１３４とサーバ１４１との間には1つ以上のルータが存在しうる）に必要とされるすべてのことはＮＨＳの支援である（図7に関して説明する）。企業ルータ１１１（ＷＡＮ出口１）はＮＨＳ＋＋機能を持つ必要がないことに留意されたい。ＷＡＮ入り口および出口ルータは同じルータである必要はないことは当業者にとって明らかであろう。

【００４２】図7は図6に示したネットワーク・トポロジに関する初期化フローの例を示す。

【００４３】フロー２０１：各ＮＨＣ（１１１，１２１，１４１）はそれぞれが所有するＮＨＳ（１３１，１３３，１３４）に対するＮＢＭＡ接続設定を開始する。この例ではＮＨＳ１３１はＮＨＣ１１１に対するサーバであり、ＮＨＳ１３３はＮＨＣ＋＋１２１に対するサーバであり、ＮＨＳ１３４はＮＨＣ１４１に対するサーバであるものと仮定される。ＮＨＣおよびそれにサービスを与えるＮＨＳの総体的所在場所は本発明の原理に関係がない。

【００４４】フロー２０２：フロー２０１が完了した後、各ＮＨＣ（１１１、１２１、１４１）はそれ自身のプロトコル・アドレスおよびそれ自身のハードウエア・アドレスをそれぞれにサービスするＮＨＳ（１３１、１

３３、１３４）に登録する（ＮＨＲＰ　ＲＥＧＩＳＴ）。例えば、ＮＤ－ＣＥ１２１はサーバ・クラスタのＩＰアドレスをそれ自身のＡＴＭアドレスと共に登録する。

【００４５】フロー２０３：アドレス登録が完了すると、ＮＨＳ１３１、１３３、１３４は肯定応答をそのクライアントに送る。

【００４６】フロー２１４：ＮＤ－ＣＥ１２１はサーバ・クラスタの各ホストとのＮＢＭＡ接続を用いて受け取った最初のパケットをルートを決められた経路に送る。各ホスト毎にそれは権限ある解決要求（ＮＨＲＰ　ＲＥＳＯＬ）をそれにサービスするＮＨＳ１３３に送る。この解決要求は宛先ホストのＩＰアドレス（例えば、サーバ１４１に対しＩＰ１）を指定する。この例では初期化フロー２１４ないし２２８がクラスタの１サーバについて説明される。これらのフローは各サーバ毎に実行されなければならない。

【００４７】フロー２１５：ＮＨＳ１３３は解決要求をその隣のＮＨＳに送る。この分野で知られた手法を用いて要求は、要求されたＩＰアドレスを有するＮＨＳ、例えばＩＰ１に対してはＮＨＳ１３４、に到達する。

【００４８】フロー２１６：ＮＨＳ１３４は１４１のハードウエア・アドレスを含む解決応答を要求を出した所に送る。

【００４９】フロー２１７：この分野で知られた手法を用いて解決応答は要求を出したところ、即ちＮＤ－ＣＥ（１２１）、に到達する。

【００５０】フロー２２８：ＮＤ－ＣＥ（１２１）はサーバ１４１とのＡＴＭショートカット接続をここで作り上げる。

【００５１】図８は、サーバ１４１および１４２を含むクラスタとのＴＣＰ接続を作るためＴＣＰクライアント１０１（図６）によって送られるＩＰデータグラムの処理の例を示す。この例はＮＤ－ＦＥがローカルにサーバを選択しようとしていないときの論理フローを述べる。上に述べたように、ＮＤ－ＦＥがサーバをローカルで選択するときには、それはＮＢＭＡ接続を再使用するのでこれについてのフローは不要である。

【００５２】フロー３０１：ＴＣＰクライアント１０１はＩＰデータグラムを送って新たなＴＣＰ接続（ＴＣＰオープン接続）を要求する。データグラムの宛先ＩＰアドレスはＩＰ＿ＳＣ、即ちサーバ・クラスタのアドレスである。ソースＩＰアドレスはＴＣＰクライアントのアドレス（ＩＰ＿ＣＬ）である。ＴＣＰヘッダはソースＴＣＰポート番号（Ｐ１）および宛先ＴＣＰポート番号（Ｐ２）を含んでいる。４つの組合せ（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）（以下、“ＴＣＰ接続キー”と呼ぶ）はＴＣＰ接続を一意的に指定する。

【００５３】フロー３０２：ＩＰデータグラムはＮＤ－ＦＥ（１１１）に到達する。ＮＤ－ＦＥ（１１１）はその「ＮＤ－ＦＥキャッシュ・テーブル」を調べてＴＣＰ接続キー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）に一致する項目を探す。これは新たな接続なので、このテーブルにこのような項目は無い。ＮＤ－ＦＥ（１１１）はこのデータグラムをデフォールト・ルートの経路３０３（図６の点線で表される）に送る。ＮＤ－ＦＥ（１１１）はまたＮＤ－ＦＥキャッシュ・テーブルに新たなＴＣＰ接続キーに対する新たな項目を作る。このＴＣＰ接続にはＮＢＭＡ接続は関連されない。

【００５４】フロー３０３：ＩＰデータグラムはルートされた経路（１３１、１３２および１３３）に沿ってすべてのルータにより送られる。

【００５５】フロー３０４：ＮＤ－ＣＥ（１２１）はＩＰデータグラムを受け取り、そのキャッシュ・テーブルを調べてＴＣＰ接続キー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）に一致する項目を探す。これは新たな接続なので、このテーブルにこのような項目は無い。要求されたサービス（ＴＣＰ宛先ポートＰ２によって指示される）、およびサーバの負荷またはその他の情報に基づいて、ＮＤ－ＣＥ（１２１）は新たなＴＣＰ接続のためのクラスタ中の最良のサーバを決定する。選択されたサーバはサーバ１４１であるものとここでは仮定される。従って、ＮＤ－ＣＥ（１２１）は以前に作られたＮＢＭＡ接続を通してＩＰデータグラムをサーバ１４１に送る（フロー２２８）。新たなＴＣＰ接続がサーバに対して作られると，ＮＤ－ＣＥ（１２１）はそれ自身のＴＣＰ接続キャッシュ・テーブルに新たな項目を加えて不活動タイマをスタートさせる。

【００５６】フロー３１５：サーバ１４１との新たなＴＣＰ接続が作られようとしているので、ＮＤ－ＣＥ１２１は修正されたＮＨＲＰ　ＲＥＧＩＳＴＥＲ要求を用いてこの新たなＴＣＰ接続をそれにサービスするＮＨＳ１３３に登録する。この修正された要求はＮＤ－ＣＥに特有の拡張フィールドを持っており、これはＴＣＰ接続キー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）、および宛先サーバのハードウエア・アドレス、即ちサーバ１４１のＡＴＭアドレスを指定する。

【００５７】フロー３１６：アドレス登録が完了すると、ＮＨＳ１３３は肯定応答をそのクライアントに送る。

【００５８】フロー３２７：短い遅延の後、ＮＤ－ＦＥ（１１１）は修正された権限ＮＨＲＰ　ＲＥＳＯＬＵＴＩＯＮ要求をそれにサービスするＮＨＳ（１３１）に送る。この修正された要求は、ＴＣＰ接続キー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）を指定するネットワーク・ディスパッチャ特有の拡張フィールドを含むことが望ましい。ＮＤ－ＦＥはルート指定された経路を用いてこの要求が満足されるまでクラスタにパケットを送り続けることになる（フロー３０３）。この要求に対して否定応答があると、ＮＤ－ＦＥ１１１は再び要求する。

【００５９】フロー３２８：この要求は権限あるものである。この分野で知られた手法を用いてこの要求はＮＢＭＡネットワーク１６２（図６）を通して送られ、要求されたＴＣＰ接続キーを有するＮＨＳ１３３に到達する。

【００６０】フロー３２９：ＮＨＳ１３３は修正されたＮＨＲＰキャッシュにおいてキー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）を用いてＴＣＰ接続キーを探索する。ＮＨＳ１３３が項目を見つけると、それは選択されたサーバ１４１のＡＴＭアドレスを指定するＮＨＲＰ　ＲＥＳＯＬＵＴＩＯＮ肯定応答を送り返す。ＮＨＳ１３３が項目を見つけないときには、それはある時間の間遅延してＮＨＣ＋＋／ＮＤ－ＣＥ１２１がＮＨＲＰ　ＲＥＧＩＳＴＥＲを送ることを許容する（フロー３１５）。遅延が終了する前にＮＨＲＰ　ＲＥＧＩＳＴＥＲが受け取られると、ＮＨＲＰ　ＲＥＳＯＬＵＴＩＯＮ肯定応答が送られ、そうでない場合にはＮＨＳ＋＋１３３がＮＨＲＰ　ＲＥＳＯＬＵＴＩＯＮ否定応答を要求に対して送る。

【００６１】フロー３３０：この分野で知られた手法を用いて、ＮＨＲＰ　ＲＥＳＯＬＵＴＩＯＮ応答は要求者であるＮＤ－ＦＥ１１１に到達する。

【００６２】フロー３４１：ＮＤ－ＦＥ（１１１）はＴＣＰ接続キー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）と共にサーバ１４１のＡＴＭアドレスをＮＤ－ＦＥキャッシュ・テーブルに保存し、サーバ１４１へのショートカットＮＢＭＡ接続を作る。ＡＴＭ接続ができると、それはそのインタフェース番号および標準的なＡＴＭのＶＰＩ／ＶＣＩ（Virtual Path Indicator/Virtual Channel Indicator）値をＮＤ－ＦＥキャッシュ・テーブルに保存する。ＮＢＭＡ接続が作り上げられる前にＴＣＰ接続に対する第２のＩＰデータグラムがＮＤ－ＦＥ１１１によって受け取られると、このデータグラムはルート決定された経路に送り出される。

【００６３】図９はＮＤ－ＦＥ１１１によるショートカットＮＢＭＡ接続の使用の例を示す。

【００６４】フロー４０１：ＴＣＰクライアント１０１は以前に作られたＴＣＰ接続（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）上のサーバ・クラスタにＩＰデータグラムを送る。

【００６５】フロー４０２：ＮＤ－ＦＥ１１１はＩＰデータグラムを受け取る。ＮＤ－ＦＥ１１１はそのＮＤ－ＦＥキャッシュ・テーブルを調べてＴＣＰ接続キー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）に一致する項目を探索し、そしてショートカットＡＴＭ接続が既に存在することを見いだす。ＮＤ－ＦＥ１１１はＩＰデータグラムをサーバ１４１に直接送る。

【００６６】フロー４１０：ＮＤ－ＦＥ１１１はルート決定された経路を通してＮＤ－ＣＥ１２１のために意図されたリフレッシュ・メッセージを定期的に送る。このメッセージは活動中のＴＣＰ接続キーのリストを含んでいる。リフレッシュ・メッセージは肯定応答を必要としない無接続データグラムであることが望ましい（例えばＵＤＰ（User Datagram Protocol）を用いて）。リフレッシュの期間はルートされるトラフィックを過剰に増大させないように十分大きく選ぶことができる。事実、ＮＤ－ＣＥ不活動タイマの期間の１／３の値で十分である。

【００６７】フロー４１１：リフレッシュ・メッセージはサーバ・クラスタ・アドレスと同じ宛先ＩＰアドレスを持っている。それはこの分野で知られた手法を用いてＮＤ－ＣＥ１２１に到達する。

【００６８】図１０はＴＣＰ接続が閉じられようとしていることを示すＴＣＰパケットの処理の例を示す。

【００６９】フロー５０１：ＴＣＰクライアント１０１はＴＣＰ接続終了を示すＴＣＰパケットを送る。ＮＤ－ＦＥ１１１はこのパケットを受け取り、ＴＣＰ接続キー（ＩＰ＿ＳＣ，ＩＰ＿ＣＬ，Ｐ２，Ｐ１）を取り出してＮＤ－ＦＥキャッシュ・テーブルを調べ、ＴＣＰ接続キーに一致する項目を探索する。それは接続状態を“クローズ”とマークし、ルート決定された経路にこのパケットを送り出し、そしてタイマを始動させる。このタイマは、最後のパケットが接続を通った後に、項目を追い出す前にどのくらいの長さだけ待つのかを指定する。この接続に対する以後のパケットはすべてこのルート指定された経路に送り出されるのでＮＤ－ＣＥ１２１は同じタイマを維持することができる。正しい動作のためには、このタイマはＭＳＬ（Maximum Segment Lifetime）の２倍よりも大きくなければならない。指示された長さに時間の間接続がアイドルになっていると、ＮＤ－ＦＥ１１１は対応する項目をそのＮＤ－ＦＥキャッシュ・テーブルから取り除く。

【００７０】フロー５０２：ＴＣＰパケット宛先アドレスはサーバ・クラスタ・アドレスであるＩＰ＿ＳＣである。従って、このパケットはそれがＮＤ－ＣＥ１２１に受け取られるまでルータからルータへとルートされる。ＮＤ－ＣＥ１２１はＴＣＰ接続キーを取り出して対応するサーバ１４１のアドレスを得る。

【００７１】フロー５０３：ＮＤ－ＣＥ１２１は以前に作られたＡＴＭ接続（フロー２２８）を通してＴＣＰパケットをサーバ１４１に送る。それはＴＣＰ接続の状態を“クローズ”とマークしてタイマを始動させる。このタイマもＭＳＬ　ＴＣＰタイマの２倍より長い。

【００７２】フロー５１０：タイマがタイムアウトすると、ＮＤ－ＣＥ１２１はそのテーブルから対応する項目を取り除く。それはまた修正されたＮＨＲＰ　ＰＵＲＧＥ要求（例えばＴＣＰ接続キーを含む）をそれにサービスするＮＨＳに送る。

【００７３】フロー５１１：ＮＨＳ１３３はその内部テーブルからＴＣＰ接続キーを取り除く。それはまたＮＨ

RP PURGE応答を要求元121に送ることによって応答する。

【0074】図11はサーバによるショートカットATM接続をクリアする処理の例を示す。

【0075】フロー601：ショートカットATM接続はサーバ141によってクリアされる。サーバは2つの理由により自由意志でショートカット接続をクリアすることができる。タイマの或るものはタイムアウトしているかまたはそのATMアドレスが変化しようとしている。この要求はNBMA接続に関連する会話の上で流れるすべてのパケットに、NBMA接続が選択されたサーバに対して再び作られるまで、ルート決定された経路を辿らせる。

【0076】フロー610：1つ以上のTCP接続項目が、クリアされようとしているNBMA接続に関連するND-FE111（ND-FEキャッシュ・テーブル）に存在するならば、NBMA接続が再び作られることが必要である。サーバのATMアドレスは認証される必要がある。修正された権限NHRP RESOLUTION（図8のフロー327で定義されたもの）要求がNHS131に送られる。サーバ141およびND-CE121が再初期化するための時間を与えるために、NHRP RESOLUTION応答はATM接続がクリアされた直後には送られないのが望ましく、それは短い遅延の後に送られる。

【0077】フロー611：この要求は権限あるものである。NHS131はこの解決要求をその隣のNHS132に送る。この要求はこの分野で知られた手法を用いて、要求されたTCP接続キーを有するNHS133に到達する。

【0078】フロー612：NHS133はその修正されたNHRPキャッシュにTCP接続キー（IP_SC，IP_CL，P2，P1）を見つけている。NHS133は、選択されたサーバ141のATMアドレスを指定するNHRP RESOLUTION肯定応答を送り返す。

【0079】フロー613：NHRP RESOLUTION応答はこの分野で知られた手法を用いて要求元のND-FE111に到達する。

【0080】フロー620：ND-FE111はサーバ141に対するATMショートカット接続を再び作る。

【0081】図12はネットワークによりATMショートカット接続をクリアする処理の例を示す。

【0082】フロー701：ND-FE111とサーバ141との間のショートカットATM接続はネットワークによりクリアされる。ND-FE111はこの接続をND-FEキャッシュ・テーブル中のどの項目によっても使用できないものとしてマークする。これは、この接続を用いたであろう後続のすべてのパケットを、選択されたサーバに対してNBMA接続が再び作られるまで、ルート決定された経路を通して送らせる。

【0083】フロー710：ND-FE111はサーバ141とのショートカット接続を再び作ろうと試みる。

【0084】高度の利用可能度およびフォールト・トレランス

図1は本発明の高度の利用可能度の特徴をも示す。スイッチング型のネットワーク1040の内部には1つ以上のND-CEが存在しうる。これらのCEはそれらの内部テーブルを同期状態に保つために上述のものと同じキャッシュに一貫性あるプロトコルを用いることになる。ND-CEが障害を起こすと、それを引き継いだND-CEが、新たな制御ND-CEであることをすべてのND-FEに知らせる。

【0085】ND-FEはND-CEとは無関係に障害を起こす。ND-FEが障害を起こすと、この障害を起こしたND-FEを介して接続されたクライアント（1021、1022、1031または1032）だけが影響を受ける。図1において、クライアント1021、1022はND-FE1014を介して接続され、クライアント1032、1031はND-FE1013を介して接続されている。ND-FE1013が障害を起こすと、期間1030を介して接続されたクライアント（1032、1031）だけが影響を受ける。クライアントの要求を本発明のシステムに導く期間ネットワーク（ルートを与える期間）が1つのND-FE1013だけを有しそれを介してネットワークがルートを与える場合には、そのFEの障害はこれらのクライアントを恒久的に切り離すことになる。この1点の障害から保護するために、第2のND-FEがこのルート基幹1030に取り付けられるように構成に組み込まれることができる。しかしながら、典型的には、インタネットのようなルート基幹では複数のルートが利用可能となる。クライアントが接続すると、ネットワークにおいて得られるルート情報が、ND-FE1013が障害を起こしたときに別のどのND-FE（またはルート）が選択されうるのかを決めるのに使用されうる。一次および二次ND-FE優先的選択情報を用いてこのように構成されうる。接続テーブルを活動中のND-FEと同期するよう維持するために一次ND-FEと次の2つの最も可能性あるND-FEとの間にキャッシュ一貫性プロトコルが維持されうる。キャッシュ一貫性プロトコルはTCP接続キーおよび選択されたサーバの識別子をバックアップのND-CEに送る。バックアップのND-CEがTCP接続キーを受け取ると、ショートカットが割り当てられ、その接続テーブルに項目が記入される。バックアップのND-FEと選択されたサーバとの間にショートカットが存在しない場合、ショートカットが作られる。最初のパケットが受け取られるまでこの接続を作るのを遅らせるために既に述べた定義のフローを用いることができることは当業者にとって明らかであろう。図1において、ND-

ＦＥ１０１３を一次ＮＤ－ＦＥとし、ＮＤ－ＦＥ１０１４をバックアップとして考察する。ＮＤ－ＦＥ１０１４接続テーブルはＮＤ－ＦＥ１０１３（１０１４はこれの代替として構成に入れられている）を介して接続されたすべてのクライアントについての項目を含んでいる。一次ＮＤ－ＦＥが障害を起こすと、基幹はパケットを構成に入れられた代替のＮＤ－ＦＥにルート指定する。ここで、代替を構成に組み入れることは、ルーティング・テーブルなどのネットワーク構成に関する従来の手法を意味する。しかしながら、その他の構成機構も利用可能であることは当業者にとって明らかであろう。例えば、ネットワーク・トポロジを動的に決定する方法が知られている。この知識は一次およびバックアップを構成するのに用いることができる。

【００８６】再び図５を参照すると、転送決定論理に関した、すべての接続、一次およびバックアップは同じ接続テーブルで与えられることが望ましい。

【００８７】一次ＮＤ－ＦＥが通常行われるように修復または回復したとき、それはルートが利用可能であるかどうかについてネットワーク１０３０を更新する。それは、それのバックアップとして構成に入れられたＦＥからキャッシュ更新を得ることになる。通常行われるように、ネットワークはルートのためのＮＤ－ＦＥの使用可能性を知るので、新たな接続が自動的にそれにルートされ、現存する接続に対するパケットはクライエントに対する中断なしにＮＤ－ＦＥを介して再ルートされることができるようになる。一次ＮＤ－ＦＥへのスイッチ・バックはこの分野で知られた標準的なプロトコルによって処理することができる。ＮＤ－ＦＥ間のキャッシュ一貫性プロトコルは本発明がパケット・スイッチング型のネットワークの利点を利用することを可能にする。

【００８８】ＮＤ－ＦＥ間のキャッシュ一貫性プロトコルは活動中の接続が終了したことがバックアップであるＮＤ－ＦＥに同報されることを保証する。輻輳状態の変動のためにネットワークがパケットを異なるやり方でルートするならば、２つのＮＤ－ＦＥに１つのクライアントが活動しているものと映ることがある。或る点で接続が終了され、そしてこれらＮＤ－ＦＥの各々は他のＮＤ－ＦＥのバックアップであるため、終了はキャッシュ一貫性プロトコルを介してバックアップＮＤ－ＦＥに繰り返されることになってバックアップＦＥからクライエント接続が外されることになる。

【００８９】本発明は好適な実施例によりその代替実施例と共に説明されたので、本明細書の特許請求の範囲に含まれる種々の等価物、改良、改善が現在、および将来においても当業者にとって自明であることが理解されよう。従って、特許請求の範囲は最初に開示された本発明の適正な保護を全うするように解釈されなければならない。

【図面の簡単な説明】

【図１】　スイッチング型のネットワーク配置された本発明の１例を示す図。

【図２】　スイッチング経路を設定するための論理的フローの１例を示す図。

【図３】　単一のスイッチに併合された制御エンジン（ＣＥ）および転送エンジン（ＦＥ）の例を示す図。

【図４】　クライアントに併合されたＦＥの例を示す図

【図５】　ＦＥにおける決定プロセスの例を示す図。

【図６】　ネクスト・ホップ・レゾリューション・プロトコル（ＮＨＲＰ）を用いたネットワーク・トポロジの例を示す図。

【図７】　種々の要素に対するＮＨＲＰ初期化フローの例を示す図。

【図８】　ＴＣＰ接続設定のためのクライアント・サーバ論理フローの例を示す図。

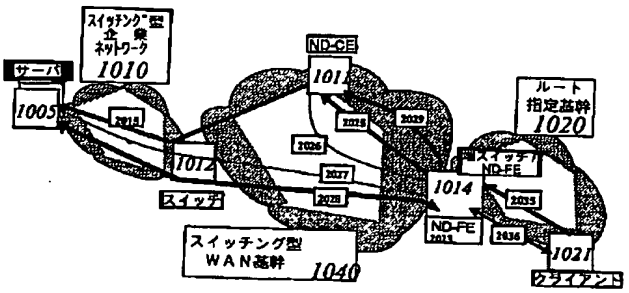
【図９】　ＴＣＰ定常状態に対するクライアント・サーバ・フローの例を示す図。

【図１０】ＴＣＰ閉鎖接続のためのクライアント・サーバ・フローの例を示す図。

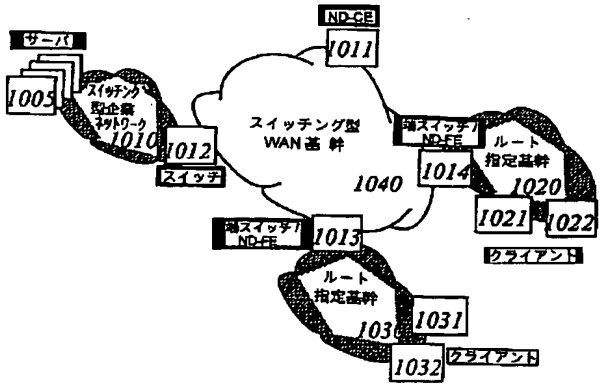
【図１１】サーバによりクリアされたクライアント・サーバ・ショートカット接続の例を示す図。

【図１２】ネットワークによりクリアされたクライアント・サーバ・ショートカット接続の例を示す図。

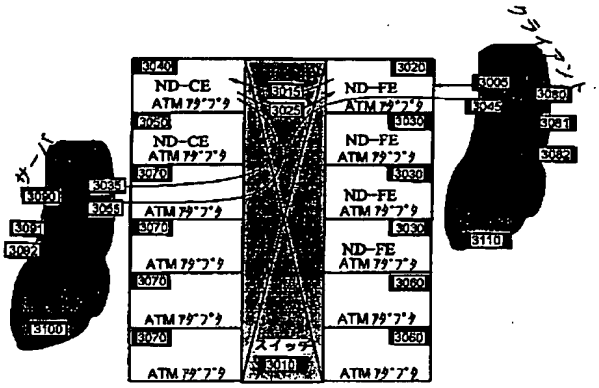
【図２】



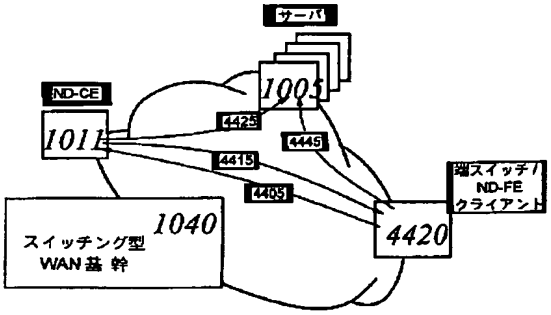
【図１】



【図３】

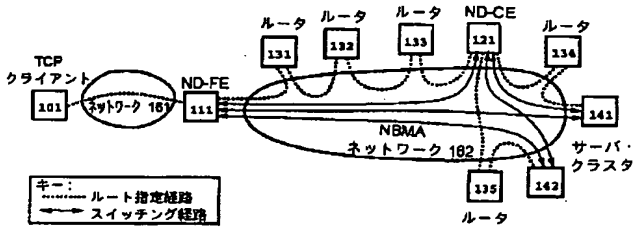


【図４】



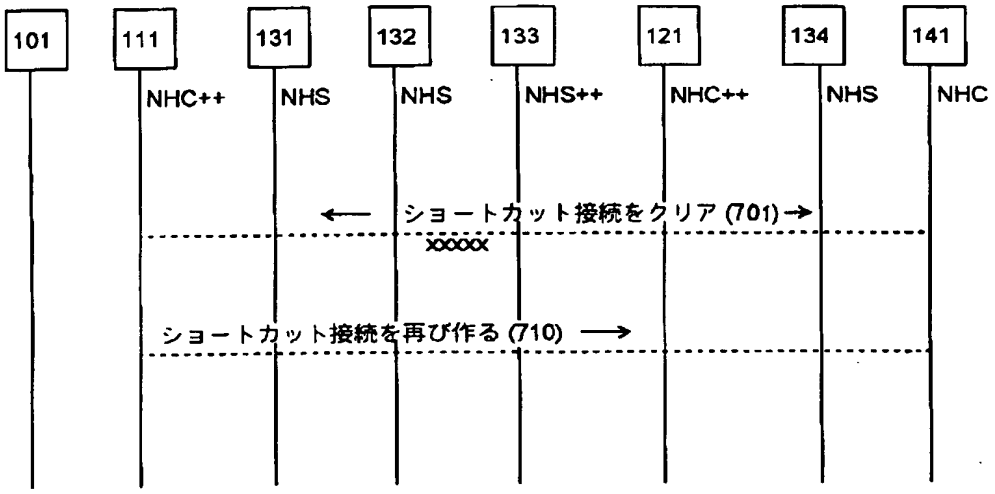
【図６】

ネットワーク・トポロジの例

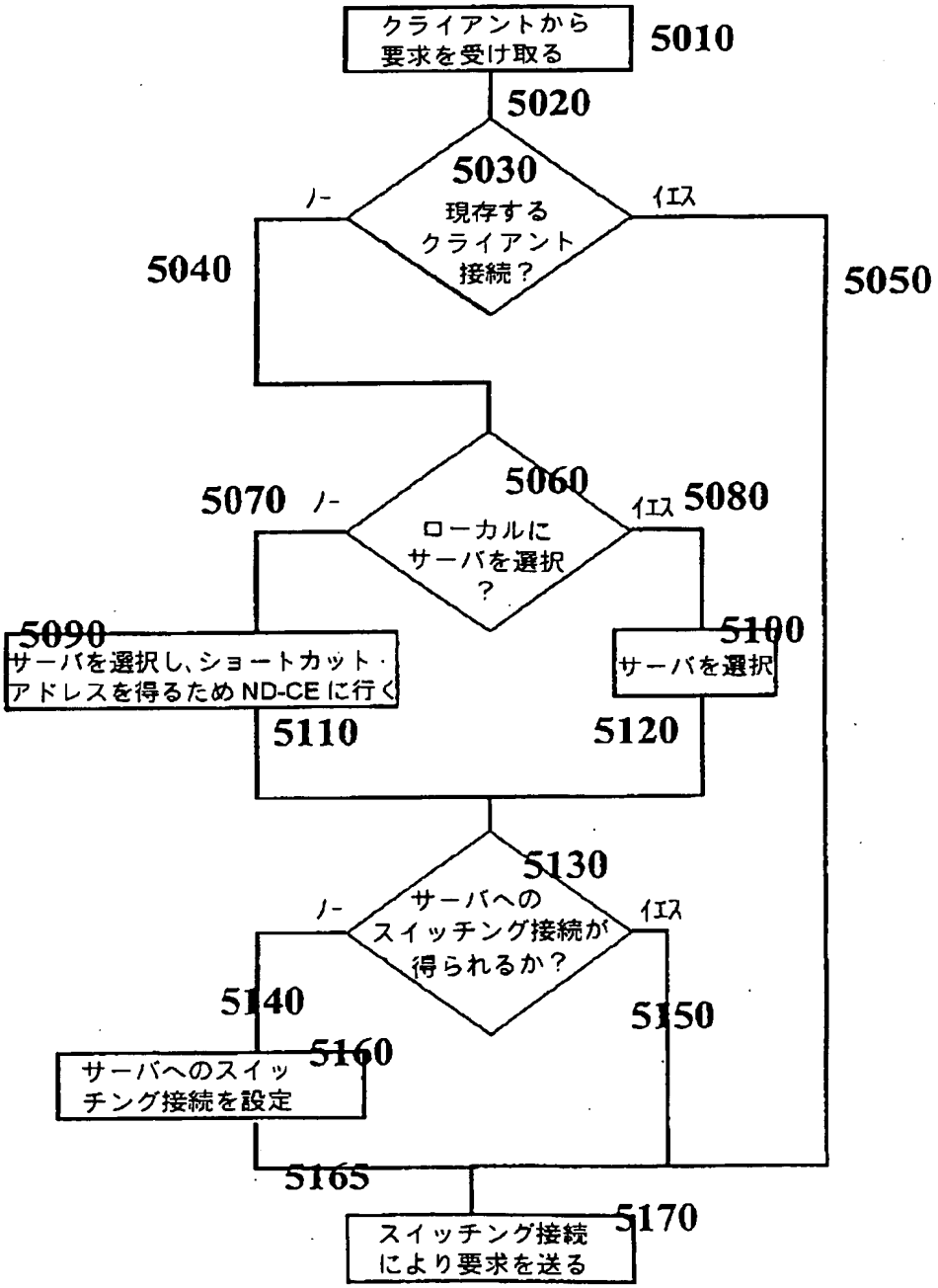


【図１２】

ＡＴＭ接続をネットワークによりクリア

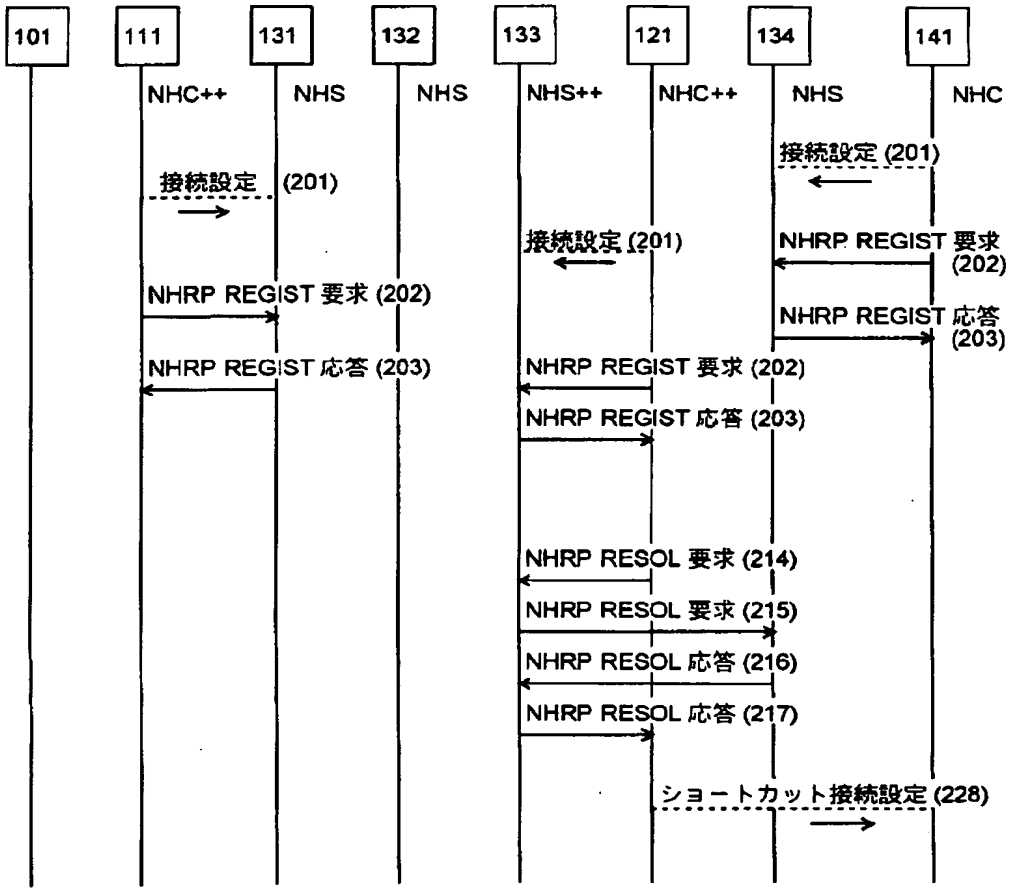


【図5】


クライアントから
要求を受け取る
5010
5020
5030
現存する
クライアント
接続？
ノー
イエス
5040
5050
5060
ローカルに
サーバを選択
？
5070
ノー
イエス
5080
5090
サーバを選択し、ショートカット・
アドレスを得るためND-CEに行く
5100
サーバを選択
5110
5120
5130
サーバへの
スイッチング接続が
得られるか？
ノー
イエス
5140
5150
5160
サーバへのスイッ
チング接続を設定
5165
5170
スイッチング接続
により要求を送る

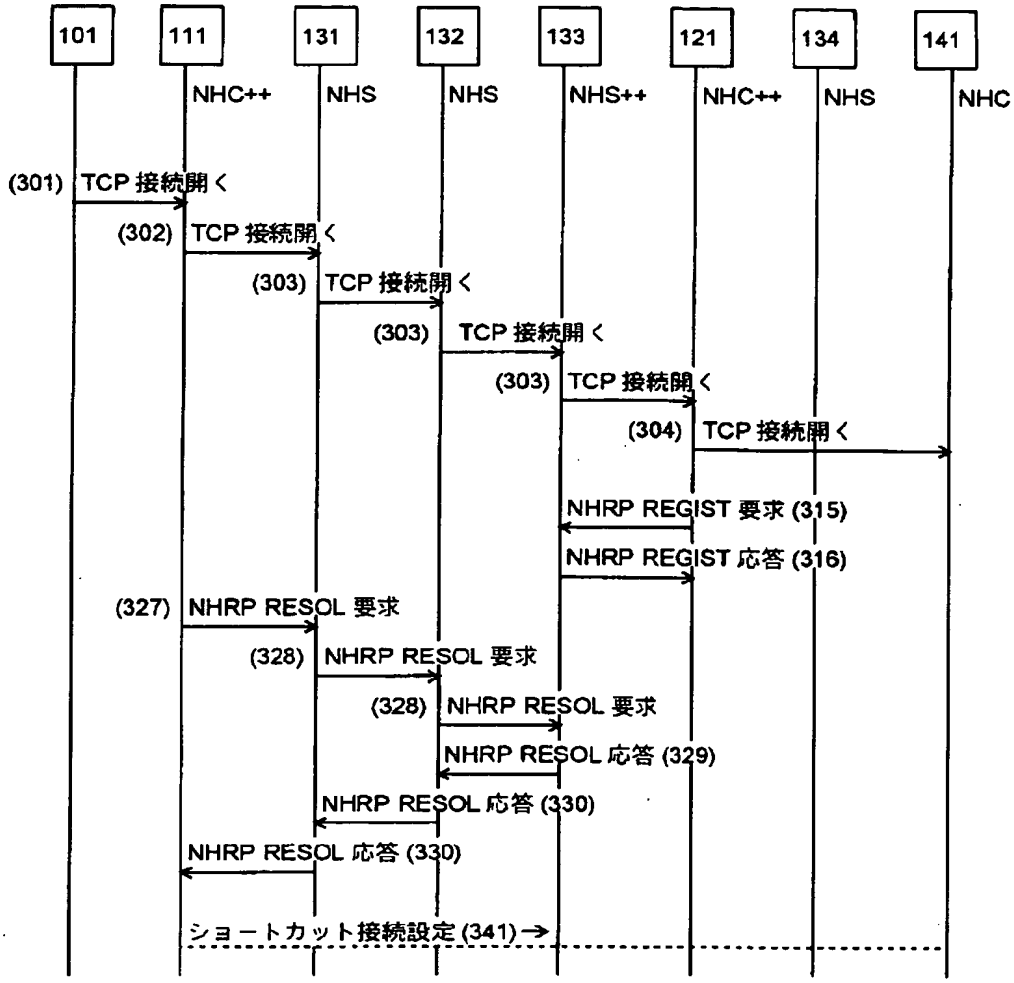
【図7】

初期化フロー


101
111
131
132
133
121
134
141
NHC++
NHS
NHS
NHS++
NHC++
NHS
NHC
接続設定 (201)
接続設定 (201)
接続設定 (201)
NHRP REGIST 要求 (202)
NHRP REGIST 要求 (202)
NHRP REGIST 応答 (203)
NHRP REGIST 応答 (203)
NHRP REGIST 要求 (202)
NHRP REGIST 応答 (203)
NHRP RESOL 要求 (214)
NHRP RESOL 要求 (215)
NHRP RESOL 応答 (216)
NHRP RESOL 応答 (217)
ショートカット接続設定 (228)

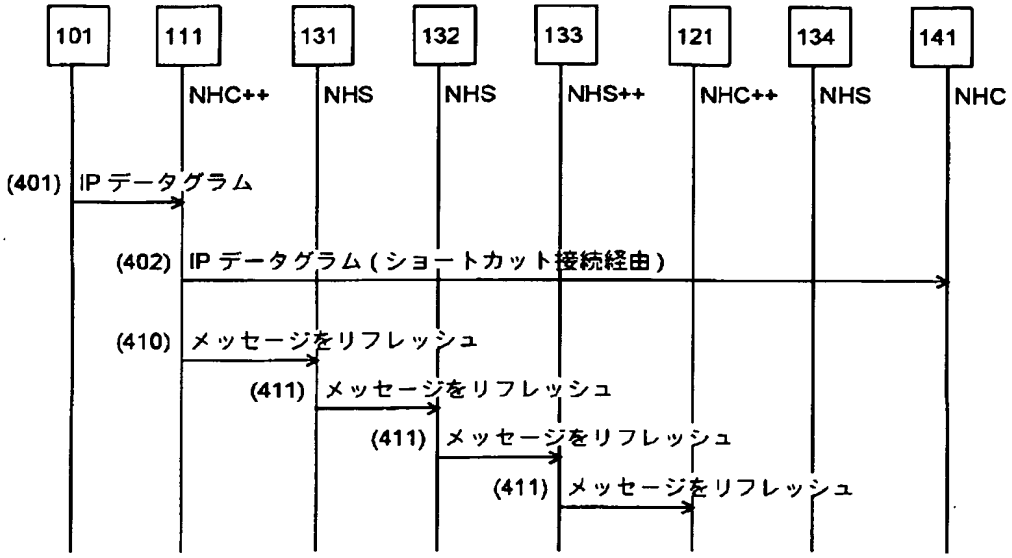
【図8】

ＴＣＰ接続設定


101
111
131
132
133
121
134
141
NHC++
NHS
NHS
NHS++
NHC++
NHS
NHC
(301) TCP 接続開く
(302) TCP 接続開く
(303) TCP 接続開く
(303) TCP 接続開く
(303) TCP 接続開く
(304) TCP 接続開く
NHRP REGIST 要求 (315)
NHRP REGIST 応答 (316)
(327) NHRP RESOL 要求
(328) NHRP RESOL 要求
(328) NHRP RESOL 要求
NHRP RESOL 応答 (329)
NHRP RESOL 応答 (330)
NHRP RESOL 応答 (330)
ショートカット接続設定 (341)

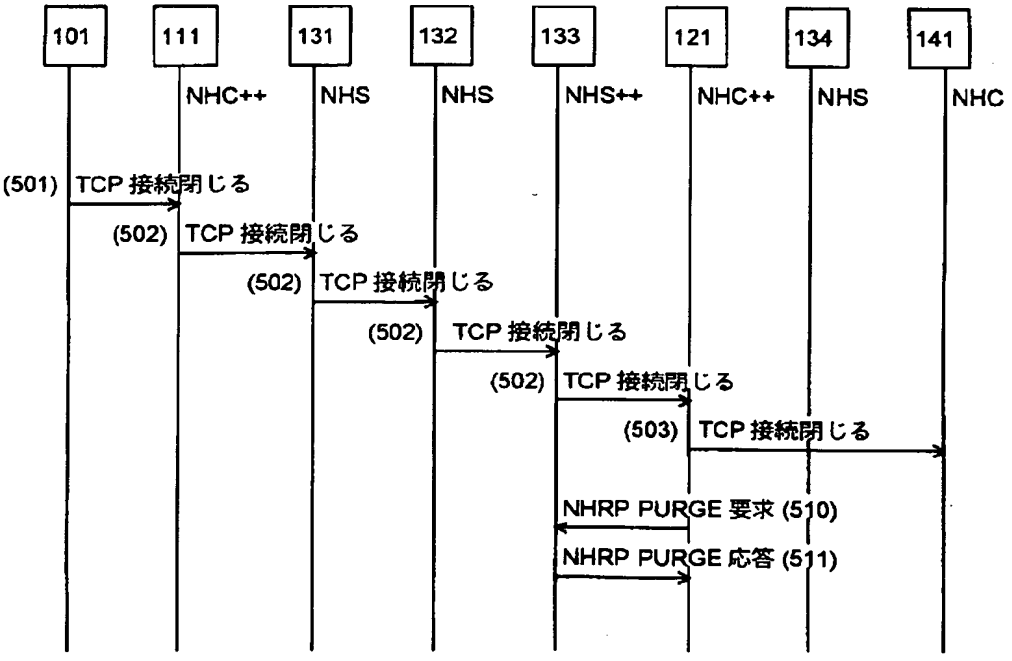
【図９】

ＴＣＰ接続ー定常状態



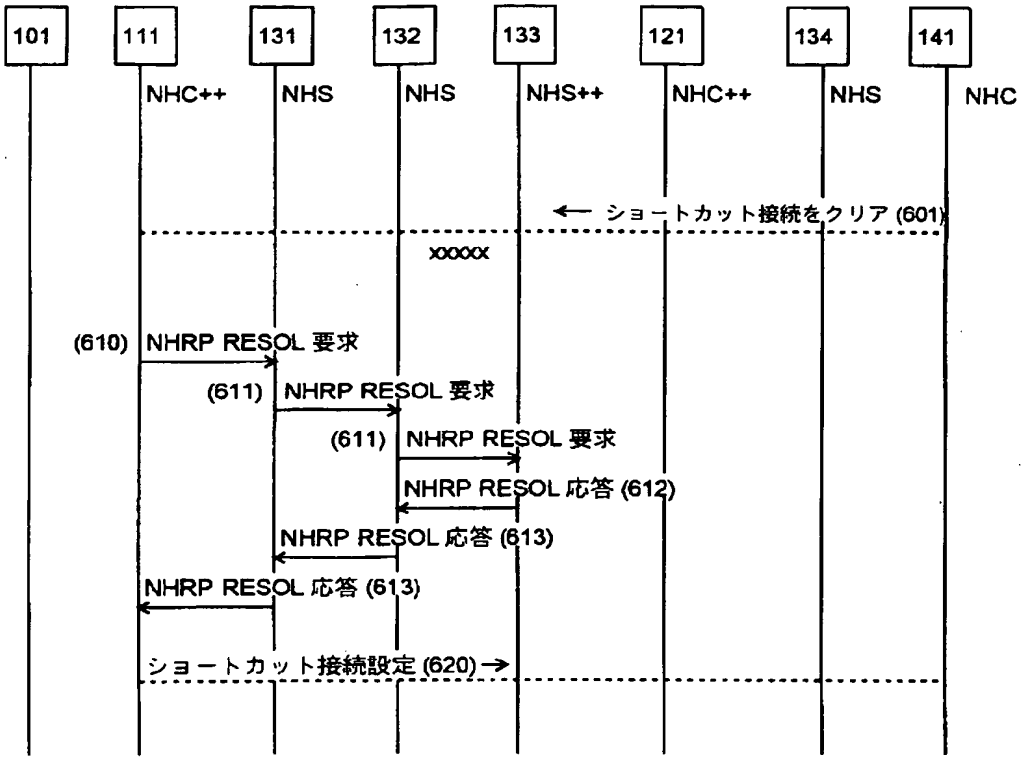
【図１０】

ＴＣＰ接続を閉じる



【図１１】

サーバによりＡＴＭ接続をクリア



フロントページの続き

(72)発明者 ガルニー・ダグラス・ホロウェイ・ハント
アメリカ合衆国10598、ニューヨーク州ヨークタウン・ハイツ、ウェリントン・コート 31

(72)発明者 エリック・ミシェル・レヴィーアベノール
フランス国06200、ニース、アンシャン・シュマン・ドゥ・ラ・ランテール 67

(72)発明者 ダニエル・ジョルジュ・ジャンーマリー・モデュ
フランス国06200、ニース、１・シュマン・ドゥ・ラ・バテリィ・リュス、ジャルダン・デ・イリ、エイ・５

Ｆターム(参考) 5K030 GA04 GA11 HA08 HC06 HC13
JT02 KA05 LB05 LB19 LE03
MB01 MD02 MD08
5K033 AA04 AA09 BA04 CB01 CB06
DA06 DB12 DB14 DB16 DB18
EA04 EB06 EC04